

DEED: A general quantization scheme for saving bits in communication

Tian Ye ^{*} Peijun Xiao [†] Ruoyu Sun [‡]

June 15, 2020

Abstract

Quantization is a popular technique to reduce communication in distributed optimization. Motivated by the classical work on inexact gradient descent (GD) [1], we provide a general convergence analysis framework for inexact GD that is tailored for quantization schemes. We also propose a quantization scheme Double Encoding and Error Diminishing (DEED). DEED can achieve small communication complexity in three settings: frequent-communication large-memory, frequent-communication small-memory, and infrequent-communication (e.g. federated learning). More specifically, in the frequent-communication large-memory setting, DEED can be easily combined with Nesterov’s method, so that the total number of bits required is $\tilde{O}(\sqrt{\kappa} \log 1/\epsilon)$, where \tilde{O} hides numerical constant and $\log \kappa$ factors. In the frequent-communication small-memory setting, DEED combined with SGD only requires $\tilde{O}(\kappa \log 1/\epsilon)$ number of bits in the interpolation regime. In the infrequent communication setting, DEED combined with Federated averaging requires a smaller total number of bits than Federated Averaging. All these algorithms converge at the same rate as their non-quantized versions, while using a smaller number of bits.

1 Introduction

There is a surge of interest in distributed learning for large-scale computation in recent decade [2–9]. In the past few years, new application scenarios such as multi-GPU computation [10–14], mobile edge computing [15, 16] and federated learning [17, 18] have received much attention. These systems are often bandwidth limited, and one important question in distributed learning is how to reduce the communication complexity.

A natural method to reduce the communication complexity is to compress the gradients transmitted between the machines. A host of recent works proposed to quantize the gradients and transfer the quantized gradients to save the communication cost [10–14]. These gradient quantization methods are successfully applied on training large-scale problems, and are shown to achieve similar performance to the original methods using less training time [13, 14]. Nevertheless, their theoretical properties, especially the relation with their un-quantized versions, are not well understood. A recent work [19] proposed DIANA and proved its linear convergence rate in a large-memory setting. Another work [20] proposed DORE which converges linearly to

^{*}Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing, Beijing (yet17@mails.tsinghua.edu.cn). The work is performed when Tian Ye is visiting UIUC.

[†]Coordinated Science Laboratory, Department of ISE, University of Illinois at Urbana-Champaign, Urbana, IL (peijunx2@illinois.edu).

[‡]Coordinated Science Laboratory, Department of ISE, University of Illinois at Urbana-Champaign, Urbana, IL (ruoyus@illinois.edu).

a bounded region in a small-memory setting. Nevertheless, these works often study a specific distribution optimization problem, and do not directly apply to other settings. For instance, they assume frequent communication, thus do not immediately apply to infrequent-communication due to the additional error caused by the infrequent updates. While it may be possible to further adapt these works to new settings, one might still wonder whether there is a unified and principled method to design quantization methods.

Our work is motivated by the classical work on inexact gradient descent (GD) [1], which provides a unified theorem for inexact GD and covers a number of algorithms (including SGD). We provide a general analysis framework for inexact GD that is tailored for quantization schemes. We also propose a quantization scheme Double Encoding and Error Diminishing (DEED). We summary our contributions below.

- **General convergence analysis.** We provide a general convergence analysis for inexact gradient descent algorithms using absolute errors in encoding. This can potentially cover a large number of quantized gradient-type methods.
- **General quantization scheme.** We propose a general quantization scheme Double Encoding and Error Diminishing (DEED). This scheme can be easily combined with existing optimization methods, and provably saves bits in communication in three common settings of distributed optimization.
- **Improved communication complexity.** In the most basic setting of large-memory frequent-communication, our theoretical bound of DEED-GD (apply DEED to GD) is at least F times better than existing works, where F is the number of bits representing a real number.

The motivation and details of our general convergence analysis and general quantization scheme will be given in Section 2 and Section 3. Discussions of related work are in the appendix. We now summarize the results of our general quantization scheme in three common settings.

Frequent-communication large-memory. We propose an algorithm DEED-GD and show it converges linearly while saves communication in bits. We further combine DEED-GD with Nesterov’s momentum to obtain an accelerated version called A-DEED-GD, which achieves the state-of-art convergence rate and save the most total number of bits in communication as shown in Table 1.

Frequent-communication small-memory. We adopt the Weak Growth Condition from [21] and prove that our algorithm DEED-SGD converges to the optimal solution at a linear rate. We compute the total number of bits to achieve a certain accuracy for both our algorithm and other works. The comparison is presented in Table 2.

Infrequent-communication. We propose DEED-Fed and provide the first explicit bound on the number of bits to achieve a certain accuracy in Federate Learning under non-i.i.d assumptions. Our results can be applied to both large-memory and small memory settings and both full-participant and partial-participant settings. To our best knowledge, [22] is the only work that also provide a convergence rate for infrequent-communication setting under realistic assumptions. However, due to the limitation of their framework, they did not do quantization in broadcasting step, which results in a great waste of communication. Our algorithm could save up to FdN bits per E iterations¹.

¹Please check Table 1 for definition of F, d and N .

Algorithm	Iterations	Bits per iteration	Total bits
DEED-GD	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dN)$	$\tilde{O}(dN\kappa \log \frac{1}{\epsilon})$
A-DEED-GD	$\tilde{O}(\sqrt{\kappa} \log \frac{1}{\epsilon})$	$\tilde{O}(dN)$	$\tilde{O}(dN\sqrt{\kappa} \log \frac{1}{\epsilon})$
DIANA [19]	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dNC)$	$\tilde{O}(dNC\kappa \log \frac{1}{\epsilon})$
ADIANA [23]	$\tilde{O}(\sqrt{\kappa} \log \frac{1}{\epsilon})$	$\tilde{O}(dNC)$	$\tilde{O}(dNC\sqrt{\kappa} \log \frac{1}{\epsilon})$
DORE [20]	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dN)$	$\tilde{O}(dN\kappa \log \frac{1}{\epsilon})$
DQGD [24]	/	$\tilde{O}(dNC)$	/
QSVRG [13]	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dNC)$	$\tilde{O}(dNC\kappa \log \frac{1}{\epsilon})$

Table 1: Summary of our theoretical results in minimizing a strongly-convex function as (1) in large-memory setting with N computing nodes. We denote the condition number of the problem as κ and the problem dimension as d . $\tilde{O}(\cdot)$ omits $\log \kappa$ and constant terms. For algorithm **DIANA**, **ADIANA**, **DQGD** and **QSVRG**, because they didn't do double quantization, they can choose either broadcasting (fully connected network) or transmitting full gradient from center to workers (star network). $C = N$ in the first case, and $C = F$ otherwise, where F is the number of bits representing a real number. **DQGD** cannot converge to the optimal solution, and is not directly comparable to our result; so we use / in the iteration cell.

Algorithm	Iterations	Bits per iteration	Total bits
DEED-SGD (WGC)	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dN)$	$\tilde{O}(dN\kappa \log \frac{1}{\epsilon})$
DIANA [19]	$\tilde{O}(\frac{\kappa}{\epsilon^2})$	$\tilde{O}(dNC)$	$\tilde{O}(dNC\frac{\kappa}{\epsilon^2})$
DORE [20]	/	$\tilde{O}(dN)$	/
QSVRG [13]	$\tilde{O}(\kappa \log \frac{1}{\epsilon})$	$\tilde{O}(dNC)$	$\tilde{O}(dNC\kappa \log \frac{1}{\epsilon})$

Table 2: Summary of our theoretical results in solving problem 1 in small-memory setting. All notations are the same as table 1.

2 General Analysis Framework

2.1 Motivation

Our work is inspired by the classical work by Bertsekas and Tsitsiklis [1] which provides a general convergence analysis of inexact gradient descent methods. Since quantization also introduces error to the update direction, a natural idea is to apply the general framework of [1] to design and analyze quantization methods. However, directly applying [1] may not provide the best result, because in quantization methods, we have a rather strong control of the “error” in the algorithm. This situation is different from the worst-case or random error considered in [1]. Our idea is to develop a modified analysis framework that can accommodate the extra freedom of controlling the quantization error, so as to obtain stronger results compared to directly applying [1]. We hope that such a general framework can help us design and analyze quantization algorithms for different settings in a unified manner.

A key element in this analysis is to use absolute error in quantization. Many theoretical works on quantization methods do not explicitly consider absolute error, but focus on relative error² [13, 24, 25]. These

²The definition of quantization by absolute error α is in Definition 2.1. The definition of quantization by relative error α is similar. We just need to replace α in the RHS of the inequality in Definition 2.1 by $\frac{\alpha}{\|w\|}$.

two types of errors are equivalent in one step quantization and only differ by scaling, but not equivalent in a multi-iteration convergence analysis. In our framework, we use absolute error in quantization so that we can sum up the absolute error over iterates and control the rate.

Overview of our general analysis framework. We first discuss why we use absolute error in quantization, and then we provide a general analysis of convergence rate. Our general analysis is not limited to algorithms that perform quantization on gradient or gradient difference. Algorithms that do quantization on weights or combination of weights and gradients are also covered in our framework. In this general analysis framework, the key component is “effective error” which we define as the error occurred at the weight. For example, in frequent-communication large-memory setting, the effective error is absolute error times the learning rate. We show that an algorithm converges only if the effective error diminishes to zero. In addition, we establish the convergence rate in terms of the learning rate and the effective error. As promised, this is a general framework, so it should be applicable to new settings with similar proofs as we will show in Section 3.4.

2.2 Content of the Framework

We consider a star-network where there are N computing nodes and one central node. Suppose we want to minimize a function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ decomposed as

$$f(w) = \frac{1}{N} \sum_{i=1}^N f_i(w), \quad (1)$$

where each function f_i is held on the i -th machine (or computing node), $i = 1, \dots, N$. We assume each f_i is L -smooth and f is μ -strongly convex. We define the condition number $\kappa := \frac{L}{\mu}$. The formal definition of L -smoothness and strong convexity are standard, so are given in the appendix.

Definition 2.1. Absolute error encoding-decoding procedure. An α -encoding-scheme of a vector w consists of an encoding algorithm $E : S \times \Xi \rightarrow \mathbb{Z}^+$ and a decoding algorithm $D : \mathbb{Z}^+ \rightarrow \mathbb{R}^d$, where S is the set of vector we need to quantize, Ξ is the set of random seeds, \mathbb{Z}^+ is the set of positive integers, and \mathbb{R} stands for the real domain. We assume:

- Unbiased coding, i.e. $\mathbb{E}_\xi[D \circ E(w, \xi)] = w$.
- The absolute error is bounded by ω , i.e. $\|D \circ E(w, \xi) - w\| \leq \alpha$.

Besides, the number of bits of this procedure is $\mathbb{E}_\xi[\log_2 |E(S, \xi)|]$.

The lemma below gives an upper bound and lower bound number of the bits with given precision.

Lemma 2.2. Given a set $S = \{x \in \mathbb{R}^d \mid \|x\|_2 \leq M\}$, any (random) quantization algorithm that encoding a vector in S by absolute error σ takes at least $\lceil d \log_2 \frac{1}{\epsilon} \rceil$ (in expectation) number of bits, where $\epsilon = \frac{\sigma}{M}$. In addition, there exists a (random) algorithm that takes only $\lceil 1.05d + d \cdot \log_2 \frac{1+2\epsilon}{\epsilon} \rceil$ bits (in expectation)³.

For convenience, we define $Q(\cdot, \epsilon)$ as a coding procedure with maximal precision ϵ with corresponding encoding and decoding procedure E_ϵ and D_ϵ . The output vector $Q(w, \epsilon)$ is $D_\epsilon \circ E_\epsilon(w, \xi)$.

³This bound is pessimistic when ϵ is large. For example, when $\epsilon \geq 1$, the lower bound is 0, however, we cannot only use 0 bits because of unbiased property. In this case, we can use sparsity to get lower bits.

To derive a general analysis for quantized GD in minimizing the problem (1), we consider a general series of functions $F_t : \mathbb{R}^d \rightarrow \mathbb{R}^d$, $t \geq 0$. The definition of F depends on the specific problems. For example, in frequent-communication, $F_t(w)$ is defined as the function mapping w_t to w_{t+1} in the t -th iteration. The only assumption on F_t in our framework is F_t is a continuous function with Lipschitz constant $c_t < 1$ with the same fixed point w^* . In most cases, this can be easily derived by strongly convexity assumption.

Assumption 2.3. For $t = 1, 2, \dots$, F_t is a continuous function with Lipschitz constant $c_t < 1$ and denote w^* as the unique fixed point of all F_t .

Theorem 2.4. Suppose Assumption 2.3 holds and $\{w_t\}$ is a sequence generated by

$$w_{t+1} = F_t(w_t) + e_t, \quad (2)$$

for some chosen initial value w_0 and e_t is a zero-mean random noise depending on the (iteration) history and is bounded by α_t . Define series $C_k^2 = \sum_{i=0}^{k-1} \alpha_i^2 \prod_{j=i+1}^{k-1} c_j^2$ and $D_k^2 = \prod_{i=0}^{k-1} c_i^2$. Then we have

$$\mathbb{E} [\|w_T - w^*\|^2] \leq D_T^2 \|w_0 - w^*\|^2 + C_T^2. \quad (3)$$

In addition, there exists functions series $\{F_i\}_{i \geq 0}$ and noise $\{e_t\}_{t \geq 0}$ to make the inequality hold. Besides, if we suppose the sequence of the Lipschitz constants $\{c_i\}$ is non-decreasing, then the right hand side of (7) converges linearly if and only if all c_t 's are always bounded above by a constant $c < 1$ and α_t converges to 0 linearly.

Remark. We leave the deterministic version of Theorem 2.4 in the appendix. It can be useful in proving the convergence of deterministic algorithm.

According to Theorem 2.4, to make $\{w_t\}$ converge to w^* , we need both D_k and C_k converge to 0. In frequent communication setting, the $D_k \rightarrow 0$ implies the summation of learning rate diverges. Then $C_k \rightarrow 0$ implies the effective error converges to 0.

The last statement of theorem 2.4 implies that for any quantized GD algorithms under our framework, we should take constant learning rate and linearly decreasing absolute error for linear convergence.

3 Application of DEED in Three Settings

Based on Theorem 2.4, we notice that using diminishing error in each iteration can guarantee fast convergence. However, according to lemma 2.2, the maximal norm of the vector we want to quantize should also be diminishing, otherwise the number of bits may explode. To avoid explosion, we choose to quantize on gradient difference instead of gradient. The intuition is that $\|\nabla f_i(w_{t+1}) - \nabla f_i(w_t)\| \leq L\|w_{t+1} - w_t\|$ who goes to zero as the iterate sequence converges. Finally, to save the communication in broadcasting, we perform quantization both on the computation nodes and the center node, i.e. “double encoding”. We name our general quantization scheme as Double Encoding and Error Diminishing (DEED).

Based on the general quantization scheme DEED, we introduce algorithms for three common settings in distributed optimization for Problem (1): frequent-communication large-memory, frequent-communication small-memory, and infrequent-communication. Frequent-communication means the every computing node communicates with the center node after every update, while this is not the case in infrequent-communication. In large-memory setting, each local server f_i has enough memory to hold its data and use them to compute the full-batch gradient of f_i . In limited memory setting (e.g. only one GPU is available in computing) that each server is only able to compute the stochastic gradients since the data cannot be fit into one server.

3.1 Frequent-communication large-memory setting

We distinguish the large-memory setting and small-memory setting for the following reasons.

First, from a practical side, different system designers have different memory budget. Some big companies can perform computation using 10,000+ GPUs or CPUs (e.g. [26–28]), while most researchers and companies can only use few GPUs or a moderate number of CPUs. The problems they are facing are indeed different, since in large-memory setting we can implement full-batch GD ⁴ (or large-batch SGD which are quite close to full-batch GD). Note that “large” is a relative term; if the system designer has only 10 CPUs or even 2 GPUs, but all data can be loaded into the memory of these machines, then this is also a large-memory setting. In a small-memory setting, we can only load a mini-batch of the dataset into one machine at a time. This necessitates the usage of Stochastic Gradient Descent (SGD).

Second, from a theoretical side, quantized gradient methods should be no better than gradient methods that utilize infinite-bandwidth. To judge the performance of quantized gradient methods, one useful metric is the gap between quantized methods and their non-quantized counter-parts. It is impossible to prove linear convergence of quantized SGD in the limited-memory setting without further assumptions, since even with infinite bandwidth SGD cannot achieve linear convergence rate [29]. In contrast, with infinite bandwidth GD can achieve linear convergence rate. Due to different upper limits, large-memory and small-memory settings should be treated separately.

The frequent-communication large-memory version of DEED is described in Algorithm 1.

Algorithm 1: Double Encoding and Error Diminishing Gradient Descent (DEED-GD)

Initialization: Each server $i \in [N]$ holds $w_0 = s_{-1}^i = v_{-1} = 0$, server 0 holds $v_{-1} = 0$, $k = 0$;

Hyper-parameters: $\eta \in (0, \frac{2}{L+\mu}]$, $c = 1 - \eta\mu$, $c' \in (c, 1)$; parameter $s \in \mathbb{R}_+$;

while the precision is not enough **do**

for $i \in [N]$ **do**

 server i computes $g_k^i = \nabla f_i(w_k)$;
 server i does quantization $d_k^i = Q(g_k^i - s_{k-1}^i, \frac{sc'^{k+1}}{2})$;
 server i updates $s_k^i = d_k^i + s_{k-1}^i$;
 server i send d_k^i to server 0;

end

server 0 computes $s_k = \frac{1}{N} \sum_{i=1}^N d_k^i + s_{k-1}$;

server 0 does quantization $u_k = Q(s_k - v_{k-1}, \frac{sc'^{k+1}}{2})$;

server 0 sends u_k to server $i, \forall i \in [N]$;

server 0 updates $v_k = u_k + v_{k-1}$;

for $i \in [N]$ **do**

 server i updates $v_k = u_k + v_{k-1}$;
 server i updates $w_{k+1} = w_k - \eta v_k$;

end

$k = k + 1$;

end

⁴Disclaimer: we discuss large-memory setting mainly due to theoretical interest. We do not run simulation on a large number of GPUs, though we will mimic the large-memory setting.

3.2 Frequent-communication small-memory

Now we consider the small-memory setting with frequent-communication. As mentioned earlier, without extra assumptions, it is impossible to prove linear convergence rate of vanilla SGD.

There are two lines of research that can prove linear convergence of SGD-type methods. Along the first line, a few variance-reduction based methods such as SVRG [30], SAGA [31] and SDCA [32] can achieve linear convergence. Along the second line, with extra assumption such as **WGC (Weak Growth Condition)**, vanilla SGD with constant stepsize can already achieve linear convergence [21]. This line of research is strongly motivated by the interpolation assumption in machine learning that the learner can fit the data, which is considered a reasonable assumption in recent literature (e.g. [21, 33]). Therefore, we focus on designing quantization algorithms along the second line.

Assumption 3.1. (*WGC Assumption [21]*) Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$, $f(x) = \frac{1}{N} \sum_{i=1}^N f_i(w)$ is the objective function. Stochastic “functions”(algorithms) $\{\bar{\nabla}_i\}_{i \in [N]}$ satisfy **WGC** if 1) $\mathbb{E}[\bar{\nabla}_i(f_i, w)] = \nabla f_i(w)$, $\forall i \in [N], w \in \mathbb{R}^d$; 2) $\frac{1}{N} \sum_{i=1}^N \mathbb{E}_{\bar{\nabla}_i} [\|\bar{\nabla}_i f_i(w)\|^2] \leq 2\rho L(f(w) - f(w^*))$.

To adapt the frequent-communication small-memory, we introduce DEED-SGD. The only differences between DEED-GD and DEED-SGD are 1) we use $\bar{\nabla}_i$ instead of the accurate gradient; 2) we use different quantization level. The full description of DEED-SGD is given in the appendix.

3.3 Infrequent-communication

A main area in distributed optimization with infrequent-communication is Federated Learning (FL), which involves training models over remote devices or data centers, such as mobile phones or hospitals, and keeping the data localized due to privacy concern or communication efficiency [17, 34]. In FL, some computation nodes might not have full participation in the updates and the data sets are non-iid. We remark that infrequent-communication is a generic design choice, and can be used in a data-center setting as well. Although existing works like QSGD do not explore this degree of freedom, infrequent communication can be combined with QSGD as well. Nevertheless, the theoretical benefit of the combination was not understood before (partially because the total number of bits was not a focus of previous works, and linear convergence rate was derived only recently [19]).

A classical algorithm in FL is Federated Averaging algorithm (FedAvg) which performs local stochastic gradient descent on computation nodes for every E iterations with a server that performs model averaging [15]. Although there have been much efforts developing convergence guarantees for FedAvg, [35–41], there is relatively scarce theoretical results on the combination of FedAvg and quantization [22, 38]. [22, 38] either make unrealistic assumptions or only perform quantization on computation nodes, and thus they are not efficient as our double encoding scheme.

We propose an algorithm called DEED-Fed. The difference between DEED-GD and DEED-Fed is that in DEED-Fed, the maximal error at iteration k is proportional to learning rate η_k . Due to space limitation, a detailed comparison between the three proposed algorithms are given in the appendix.

3.4 Theoretical Analysis

In this section, we give the computational and communication complexity of the algorithms DEED-GD, DEED-SGD and DEED-Fed. Since all of them are in the same framework, their proofs and results are similar.

We put it into one single theorem.

Theorem 3.2. *Consider solving Problem (1) under one of the three settings by the corresponding algorithms (DEED-GD, DEED-SGD or DEED-Fed). Assume f_i is L -smooth and f is μ convex. Assume all f_i 's are μ convex in **DEED-Fed**. Denote w_t as the iterate at iteration t and w^* is the optimal solution of Problem (1).*

For DEED-GD, we choose the learning rate $\eta_t \equiv \eta = \frac{2}{L+\mu}$, $c := 1 - \eta\mu$, $c < c' < 1$, and the maximal error at iteration t is $sc'^{t+1}/2$ where s is the quantization level.

For DEED-SGD, we assume (Weakth Growth Condition) WGC is satisfied for approximate gradient $\bar{\nabla}_i$ for every f_i with parameter ρ . We choose the learning rate $\eta_t \equiv \eta = \frac{1}{\rho L}$, $c := 1 - \eta\mu$, $c < c' < 1$, and the maximal error at iteration t is $\sqrt{sc'^{k+1}}/2$ and the error is unbiased.

For DEED-Fed, we choose the learning rate $\eta_t := \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$, $\gamma > 1$ such that $\eta_0 \leq \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. Let the maximal error at iteration $t \in \{0, E, 2E, \dots\}$ be $s\eta_t$.

We have the following results:

- DEED-GD communicates $\tilde{O}(Nd)$ bits at iteration $t \geq 1$, and

$$\|w_t - w^*\| \leq (c')^t \left(\max \left\{ 0, \|w_0 - w^*\| - \frac{c\eta s}{c' - c} \right\} + \frac{c'\eta s}{c' - c} \right).$$

- DEED-SGD communicates $\tilde{O}(Nd)$ bits at iteration $t \geq 1$, and

$$\mathbb{E}\|w_t - w^*\|^2 \leq (c')^t \left(\max \left\{ 0, \|w_0 - w^*\|^2 - \frac{c\eta^2 s}{c' - c} \right\} + \frac{c'\eta^2 s}{c' - c} \right).$$

- DEED-Fed communicates $\tilde{O}(Nd)$ bits at iteration $t \in \{E, 2E, \dots\}$, and

$$\mathbb{E}\|\bar{w}_t - w^*\|^2 \leq \frac{v}{\gamma + t},$$

where v is some constant dependent on the Federated learning settings (e.g. full participant or not) as well as the the initial error $\|w_0 - w^*\|^2$.

Remark 1: Based on these results, we can easily compute the total number of bits needed to achieve a certain accuracy; see Table 1 and Table 2.

Remark 2: Our result allows to trade-off communication time and computation time. By changing the parameters c' and s , we can find optimal choice of convergence speed and error size.

Theorem 3.2 implies that to achieve $\|w_T - w^*\| \leq \varepsilon$, we need $\tilde{O}(\kappa \log \frac{1}{\varepsilon})$ iterations for DEED-GD and DEED-SGD and $O(\frac{1}{\varepsilon^2})$ iterations for DEED-Fed. These convergence rates match those of the corresponding algorithms with infinite bandwidth. Due to space limitation, we eliminate the detailed definitions of some constants in Theorem 3.2 and we will provide the details in the appendix.

4 Quantization of Nesterov acceleration

In frequent-communication large-memory setting, we combine Nesterov's acceleration with our quantization scheme DEED. The accelerating algorithm is very similar to Algorithm 1. The only difference between Algorithm 1 and this accelerated version is that we add momentum in the final update step. The full description of the algorithm is given in the appendix.

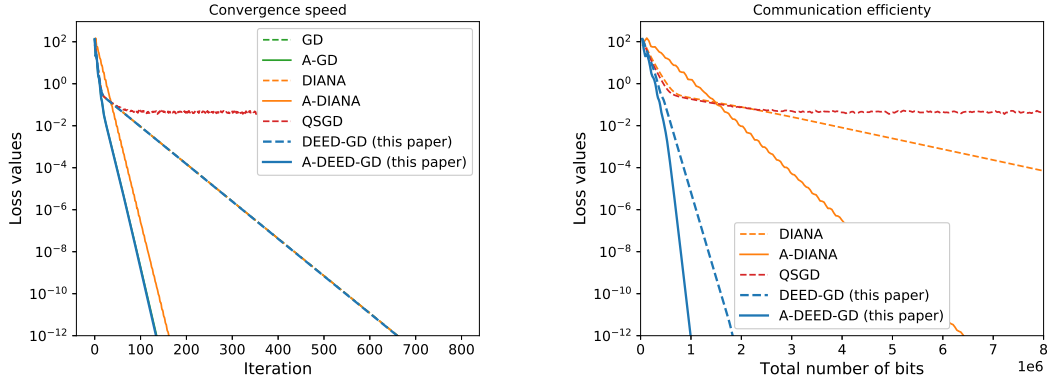


Figure 1: Compare our algorithms with other works on Linear Regression Problem.

Theorem 4.1. Consider solving Problem 1 by Algorithm **A-DEED-GD** and assume each f_i is L -smooth and f is μ -strongly convex. Let the learning rate $\eta = \frac{1}{L}$, the constant $c := \sqrt{1 - \sqrt{\frac{\mu}{L}}}$ such that $c < c' < 1$, and we do quantization with maximal error $sc'^{k+1}/2$ at every iteration k . Then we have:

- (1) $\|x_k - x^*\| \leq \sqrt{\frac{2}{\mu}} \cdot c'^k \cdot \sqrt{\Delta/\gamma^{2k} + C}$, where $C = \beta_s^2 + \alpha_s + \beta\sqrt{\alpha_s} - \Delta$, $\alpha_s = \frac{s^2}{L(\gamma^2-1)} + \Delta$, $\beta_s = \left(\frac{3\sqrt{\frac{2}{L}} + 5\sqrt{\frac{2}{\mu}}}{c(\gamma^2-1)} \right) s\gamma$, $\gamma = c'/c > 1$.
- (2) The number of bits at iteration $k \geq 1$ is $O(Nd)$.

Theorem 4.1 implies that we can improve the linear convergence rate in frequent-communication large-memory setting from $O(\kappa)$ to $O(\sqrt{\kappa})$ with acceleration trick, where κ is the condition number of the objective function. This also provides a fewer total number of bits in communication.

Remark 1: We separate algorithm **A-DEED-GD** out from previous three algorithms because we cannot directly use theorem 2.4 due to the momentum. But the intuition and technique are very similar. Hence we put it in the **DEED** series.

Remark 2: We noticed an independent work [23] which also proved an accelerated rate, but differs from our work in the following aspects. First, their encoding scheme is a non-trivial combination of the Nesterov's momentum and DIANA (which is why a separate paper is written), while our combination is rather straightforward. Second, their bound has an extra constant dependent on the communication scheme (can be N or number of encoding bits) while our bound does not. Third, our work aims to develop a general framework, and acceleration is just one case; while their work focused on acceleration in large-memory frequent-communication setting.

5 Experiment

Linear regression. We empirically validate our approach on linear regression problem as shown in Figure 1. The solid lines correspond to gradient descent type of algorithms and the dashed lines correspond to the accelerated versions. In the left figure, the curves of our method coincide with the curves of the un-quantized baseline methods (GD and A-GD). QSGD performs the worst since it uses constant error. In the right

figure, our accelerated method achieves high-accuracy solution with the fewest number of bits, and another state-of-art algorithm takes more than 6 times of bits than ours to reach the same accuracy. Even without acceleration, our algorithm (DEED-GD) takes fewer bits than A-DIANA. Overall, our algorithms save the most number of bits in communication without scarifying the convergence speed.

Image Classification. We also compare our algorithm with other state-of-art algorithms (e.g. QSGD, TernGrad, DoubleSqueeze, DIANA) on image classification tasks on MNIST data set [42]. The results still show that our algorithms outperform others both in terms of convergence speed and communication complexity. The details of the two experiments are provided in the appendix.

6 Conclusion

In this paper, we provide a general convergence analysis for inexact gradient descent algorithms using absolute errors, that is tailored for quantized gradient methods. Using this general convergence analysis, we derive a quantization scheme named DEED and propose algorithms for three common settings in distributed optimization: frequent-communication large-memory, frequent-communication small-memory, and infrequent-communication (both large-memory and small-memory included). We also combine DEED with Nesterov’s acceleration to provide an accelerated algorithm A-DEED-GD for frequent-communication large-memory, which improves the convergence rate from $O(\kappa)$ to $O(\sqrt{\kappa})$. Our proposed algorithms converge almost as fast as their non-quantized versions and save communication in terms of bits. We empirically test our algorithms on linear regression problems and image classification tasks, and find that they use fewer bits than other algorithms.

A Outline, Definition and Related Work

In the appendix, we first discuss some common definitions and related work. In Section B, we provide the proof appeared in the general convergence analysis in Section 2. In Section C, we give the detailed description of the three algorithms based on our framework DEED. In Section D, we provide the proof of the theorems for DEED in the three settings. In Section F, we illustrate the efficiency of our algorithms in linear regression problem and image classification tasks on MNIST dataest.

Definition A.1. A differentiable function $h : \mathbb{R}^d \rightarrow \mathbb{R}$ is L -smooth if $\forall x, y \in \mathbb{R}^d$,

$$|h(y) - h(x) - \langle y - x, \nabla h(x) \rangle| \leq \frac{L}{2} \|x - y\|^2 \quad (4)$$

Definition A.2. A differentiable function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -strongly-convex if $\forall x, y \in \mathbb{R}^d$,

$$f(y) - f(x) - \langle y - x, \nabla f(x) \rangle \geq \frac{\mu}{2} \|x - y\|^2 \quad (5)$$

Definition A.3. Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}, f(w) := \frac{1}{N} \sum_{i=1}^N f_i(w)$, where f_i is L -smooth $\forall i \in [N]$ and f is μ -strongly-convex. Then the condition number κ of this collection of functions is defined as $\kappa = \frac{L}{\mu}$.

Assumption A.4. Assume that in (1) f is μ -strongly-convex and f_i is L -smooth $\forall i \in [N]$.

Related work. The study of the communication complexity in terms of bits for convex minimization of the problem (1) can be traced back to a classical work [43] in 1987. This work focuses on the two-nodes case for frequent-communication large-memory setting, and proposed a nearly optimal algorithm using quantized gradient differences. For multiple-nodes cases, [19] provides a linear convergence rate also using gradient differences on computing nodes. In frequent-communication small-memory setting, to save more bits in communication, [24, 25] consider double encoding on gradient differences. In infrequent-communication setting, [22] uses quantized gradient differences and proves sublinear convergence rate. Our work is different, as we provide convergence analysis on three settings for multiple-node cases and use doubling encoding to save more bits.

B Proof of Theorems for General Convergence Analysis

Lemma B.1. Given a set $S = \{x \in \mathbb{R}^d \mid \|w\|_2 \leq M\}$, any (random) quantization algorithm that encoding a vector in S by absolute error σ takes at least $\lceil d \log_2 \frac{1}{\epsilon} \rceil$ (in expectation) number of bits, where $\epsilon = \frac{\sigma}{M}$. In addition, there exists a (random) algorithm that takes only $\lceil 1.05d + d \cdot \log_2 \frac{1+2\epsilon}{\epsilon} \rceil$ bits (in expectation).

Proof sketch. For the lower bound, we only need to prove the deterministic version since every random algorithm can be reduced to a deterministic algorithm by fixing ξ . Then it is equivalent to cover S with small balls and proof follows. And we will use constructive method to prove the upper bound.

Proof. We first prove the lower bound. $\forall m \in \mathbf{E}(S)$, construct a ball centered at $\mathbf{D}(m)$ with radius σ . Then all these balls form a cover of S . Otherwise there is a vector $v \in S$ outside the cover and

$$\|v - \mathbf{D}(\mathbf{E}(v))\|_2 \geq \min_{m \in \mathbf{E}(S)} \|v - \mathbf{D}(m)\|_2 > \sigma,$$

which contradicts to the assumption.

Hence, the sum of the volumes of these small balls is not less than the volume of S . Finally, $|\mathbf{E}(S)| \geq \frac{M^d}{\sigma^d}$ follows.

On the other hand, we divide the whole space by cubes with side length $\frac{2\sigma}{\sqrt{d}}$ regularly. Then select the cubes who have non-empty intersection with S . Then every point in S is contained in a cube, which means it can be encoded (unbiasedly) by the vertices of the cube with maximal error σ . Then these cubes must be contained in ball $B(0, M + 2\sigma)$ since the diameter of the cube is 2σ . In this case, the number of cubes is at most $\frac{\frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)}(M+2\sigma)^d}{\left(\frac{2\sigma}{\sqrt{d}}\right)^d}$, and then the number of bits is at most $\left\lceil \log_2 \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \left(\frac{(1+2\varepsilon)\sqrt{d}}{2\varepsilon}\right)^d \right\rceil$.

Recall Stirling's formula $\Gamma(n+1) \geq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n$, we have

$$\begin{aligned} \log_2 \frac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \left(\frac{(1+2\varepsilon)\sqrt{d}}{2\varepsilon}\right)^d &\leq \log_2 \frac{\pi^{\frac{d}{2}}}{\sqrt{\pi d} \left(\frac{d}{2e}\right)^{\frac{d}{2}}} \left(\frac{(1+2\varepsilon)\sqrt{d}}{2\varepsilon}\right)^d \\ &= d \cdot \log_2 \frac{\sqrt{2\pi e}}{(\pi d)^{\frac{1}{2d}}} \frac{1+2\varepsilon}{2\varepsilon} \\ &\leq d \cdot \log_2 \sqrt{\frac{\pi e}{2}} \frac{1+2\varepsilon}{\varepsilon} \\ &\leq 1.05d + d \cdot \log_2 \frac{1+2\varepsilon}{\varepsilon}. \end{aligned}$$

□

Theorem B.2. Suppose Assumption 2.3 holds and $\{w_t\}$ is a sequence generated by

$$w_{t+1} = F_t(w_t) + e_t, \quad (6)$$

for some chosen initial value w_0 and e_t is a zero-mean random noise depending on the (iteration) history and is bounded by α_t . Define series $C_k^2 = \sum_{i=0}^{k-1} \alpha_i^2 \prod_{j=i+1}^{k-1} c_j^2$ and $D_k^2 = \prod_{i=0}^{k-1} c_i^2$. Then we have

$$\mathbb{E} [\|w_T - w^*\|^2] \leq D_T^2 \|w_0 - w^*\|^2 + C_T^2. \quad (7)$$

In addition, there exists functions series $\{F_i\}_{i \geq 0}$ and noise $\{e_t\}_{t \geq 0}$ to make the inequality hold. Besides, if we suppose the sequence of the Lipschitz constants $\{c_i\}$ is non-decreasing, then the right hand side of (7) converges linearly if and only if all c_i 's are always bounded above by a constant $c < 1$ and α_t converges to 0 linearly.

Proof. The inequality (7) is straightforward.

$$\begin{aligned} \mathbb{E} [\|w_{t+1} - w^*\|^2 | w_t] &= \mathbb{E} [\|F_t(w_t) + e_t - w^*\|^2 | w_{T-1}] \\ &= \mathbb{E} [\|F_t(w_t) - w^*\|^2 | w_{T-1}] + \alpha_t^2 \\ &\leq c_t^2 \|w_t - w^*\|^2 + \alpha_t^2. \end{aligned}$$

Then we can prove inequality (7) by mathematical induction. For $T = 0$, we have $\|w_0 - w^*\|^2 \equiv \|w_0 - w^*\|^2$. Suppose it holds for $T \leq k$, we have

$$\begin{aligned} \mathbb{E} \|w_{k+1} - w^*\|^2 &\leq c_k^2 \mathbb{E} \|w_k - w^*\|^2 + \alpha_k^2 \\ &\leq c_k^2 (D_k^2 \|w_0 - w^*\|^2 + C_k^2) + \alpha_k^2 \end{aligned}$$

$$= D_{k+1}\|w_0 - w^*\|^2 + C_{k+1}^2.$$

The inductions succeed.

Next, we define $F_t(x) := c_t x$. Given history $\{w_0, w_2, \cdot, w_t\}$, we assign \bar{e}_t as an arbitrary vector orthogonal to $F_t(w_t) - w^*$ with length α_t . Define $e_t = \bar{e}_t$ with probability $\frac{1}{2}$, and $e_t = -\bar{e}_t$ otherwise. In this case, we always have

$$\|w_{t+1} - w^*\|^2 = c_t^2 \|w_t - w^*\|^2 + \alpha_t^2,$$

and

$$\|w_{k+1} - w^*\|^2 = D_{k+1}\|w_0 - w^*\|^2 + C_{k+1}^2$$

follows.

Finally, suppose the sequence of the Lipschitz constants $\{c_i\}$ is non-decreasing.

- **Necessity.** Suppose there exists constant $C, M > 0$ and $c \in (0, 1)$ such that $D_T^2\|w_0 - w^*\|^2 + C_T^2 \leq Cc^T$, $\forall T \geq M$.

We firstly prove a lemma.

Lemma B.3. The function $g(x) := \frac{\ln \frac{1}{1-x}}{x}$ is increasing on $(0, 1)$.

Proof. $\forall x \in (0, 1)$, we have $g'(x) = \frac{\frac{x}{1-x} + \frac{\ln \frac{1}{1-x}}{x^2}}{x^2} > 0$. □

Recall $c_t \geq c_0$, $\forall t \geq 0$, we have $\frac{\ln \frac{1}{c_t}}{1-c_t} \leq \frac{\ln \frac{1}{c_0}}{1-c_0}$, i.e. $\ln \frac{1}{c_t} \leq C_0(1-c_t)$ where $C_0 := \frac{\ln \frac{1}{c_0}}{1-c_0}$. Because $D_T^2 \leq \frac{C}{\|w_0 - w^*\|^2} c^T$, we have $\sum_{i=0}^{T-1} 2 \ln \frac{1}{c_i} \geq \ln \frac{\|w_0 - w^*\|^2}{C} + T \ln \frac{1}{c}$. Moreover, $2C_0 \sum_{i=0}^{T-1} (1-c_i) \geq \ln \frac{\|w_0 - w^*\|^2}{C} + T \ln \frac{1}{c}$. This suggests $\forall k \geq 0$, $\lim_{T \rightarrow \infty} \frac{\sum_{i=k}^{T-1} (1-c_i)}{T} \geq \frac{\ln \frac{1}{c}}{2C_0}$, and $1 - c_k \geq \frac{\ln \frac{1}{c}}{2C_0}$ follows. On the other hand, $Cc^T \geq C_T^2 \geq \alpha_{T-1}^2$. Hence, c_i is bounded by $c := 1 - \frac{\ln \frac{1}{c}}{2C_0}$ and α_i diminishes exponentially.

- **Sufficiency.** Suppose $c_i \leq c < 1, \forall i \geq 0$ and $\alpha_t \leq C\alpha^t, \forall t \geq M$ for constant $C > 0, M \geq 0$. Without loss of generality, we assume $M = 0$. We only need to prove D_k and C_k diminish exponentially separately. It is trivial for D_k . For C_k , we have $C_k^2 := C_k^2 = \sum_{i=0}^{k-1} \alpha_i^2 \prod_{j=i+1}^{k-1} c_j^2 \leq Ck\beta^k$ where $\beta = \max\{c, \alpha\}$. Then C_k diminishes exponentially since $Ck\beta^k \leq \beta^{\frac{k}{2}}$ for sufficient large k . □

C Algorithms

C.1 DEED-GD

The pseudo-code of DEED-GD is given in Algorithm 1.

C.2 DEED-SGD

The algorithm of DEED-SGD is given below (algorithm 2). As we promised previously, there are only two differences between DEED-GD and DEED-SGD.

- In line 5, we use approximate gradient $\bar{\nabla}_i$ instead of gradient ∇ .
- In line 2, line 6 and line 11, we use different η , c and c' . And the maximal error in quantization is changed.

Algorithm 2: Double Encoding and Error Diminishing for Stochastic Gradient Descent (DEED-SGD)

- 1: Initialization: Each server $i \in [N]$ holds $w_0 = s_{-1}^i = v_{-1} = 0$, server 0 holds $v_{-1} = 0$, $k = 0$;
 - 2: $\eta = \frac{1}{\rho L}$, $c = 1 - \frac{\mu}{\rho L}$, $c < c' < 1$ and $s \in \mathbb{R}_+$ is the quantization level;
 - 3: **while** the precision is not enough **do**
 - 4: **for** $i \in [N]$ **do**
 - 5: server i computes $g_k^i = \bar{\nabla}_i f_i(x_k)$;
 - 6: server i does quantization $d_k^i = Q(g_k^i - s_{k-1}^i, \sqrt{sc'^{k+1}}/2)$;
 - 7: server i updates $s_k^i = d_k^i + s_{k-1}^i$;
 - 8: server i send d_k^i to server 0;
 - 9: **end for**
 - 10: server 0 computes $s_k = \frac{1}{N} \sum_{i=1}^N d_k^i + s_{k-1}$;
 - 11: server 0 does quantization $u_k = Q(s_k - v_{k-1}, \sqrt{sc'^{k+1}}/2)$;
 - 12: server 0 sends u_k to server i , $\forall i \in [N]$;
 - 13: server 0 updates $v_k = u_k + v_{k-1}$;
 - 14: **for** $i \in [N]$ **do**
 - 15: server i updates $v_k = u_k + v_{k-1}$;
 - 16: server i updates $w_{k+1} = w_k - \eta v_k$;
 - 17: **end for**
 - 18: $k = k + 1$;
 - 19: **end while**
-

C.3 DEED-Fed

We first introduce the original FedAvg algorithm. Define $F(w) := \sum_{i=1}^N p_i F_i(w)$ where F_i 's are μ -convex and L -smooth functions defined on \mathbb{R}^d and $p_i \geq 0$, $\sum_{i=1}^N p_i = 1$. In each round, say round $k \geq 0$, the center server sends weight w_{tE} to N slave nodes, and the k^{th} slave nodes performs E local updates (for i in $\{0, 1, \dots, E-1\}$):

$$w_{tE+i+1}^k = w_{tE+i}^k - \eta_{tE+i} \nabla F_k(w_{tE+i}^k, \xi_{tE+i}^k).$$

Finally, in **full participant setting**, all slave nodes sends their final weights to the center server, and

center server computes

$$w_{(t+1)E} := \sum_{i=1}^N p_i w_{(t+1)E}^i.$$

For **partial participant setting**, not all slave nodes stay active in each round. Here are two more detailed settings.

- 1) In setting 1, we defined a set S_{t+1} of K indices selected randomly with replacement from $[N]$ with probability distribution (p_1, \dots, p_N) . Then the center server updates

$$w_{(t+1)E} := \frac{1}{K} \sum_{i \in S_{t+1}} w_{(t+1)E}^i. \quad (8)$$

- 2) In setting 2, we defined a set S_{t+1} of K indices selected evenly and randomly without replacement from $[N]$. Then the center server updates

$$w_{(t+1)E} := \frac{N}{K} \sum_{i \in S_{t+1}} p_i w_{(t+1)E}^i. \quad (9)$$

Except the assumption of smoothness and convexity, there are two more assumptions.

Assumption C.1. Let ξ_t^k be sampled from the k^{th} device's local data uniformly at random. The variance of stochastic gradients in each device is bounded: $\mathbb{E} \|\nabla F_k(w_t^k, \xi_t^k) - \nabla F_k(w_t^k)\|^2 \leq \sigma_k^2$.

Assumption C.2. The expected squared norm of stochastic gradients is uniformly bounded, i.e. $\mathbb{E} \|\nabla F_k(w_t^k, \xi_t^k)\|^2 \leq G^2$.

Based on these two assumptions, their theorems said

Theorem C.3. With the algorithm above, we have

$$\mathbb{E} \|\bar{w}_{t+1} - w^*\|^2 \leq (1 - \eta_t \mu) \mathbb{E} \|\bar{w}_t - w^*\|^2 + \eta_t^2 (B + C),$$

where $B = \sum_{k=1}^N p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2$, and C is a constant depending on different setting. In full participant setting, $C = 0$. In partial participant setting 1, $C = \frac{4}{K} E^2 G^2$, and in partial participant setting 2, $C = \frac{N-K}{N-1} \frac{4}{K} E^2 G^2$.

The algorithm **DEED-Fed** is a simple combination of DEED-GD and Federated Averaging algorithms. Algorithm 3 is the pseudo-code of fully-participant setting.

To change algorithm 3 into partial participant versions, we only need to replace $[N]$ by set S in line 9 and change the summation in line 14 into (8) or (9) correspondingly.

D Theorems for DEED in Three Settings

In this section, we restate the theorem 3.2 separately and give proof separately, too.

Algorithm 3: Double Encoding and Error Diminishing Federated Averaging (DEED-Fed)

```

1: Initialization: Each server  $i \in [N]$  holds  $w_0 = s_{-E}^i = v_{-E} = 0$ , server 0 holds  $v_{-E} = 0$ ,  $k = 0$ ;
2: Hyper-parameters:  $\eta_k \in \mathbb{R}_+$ , parameter  $s \in \mathbb{R}_+$ ;
3: while the precision is not enough do
4:   for  $i \in [N]$  do
5:      $w_{k+1}^i = w_k^i - \eta_k \nabla f_i(w_k^i, \xi_k^i)$ ;
6:      $k = k + 1$ ;
7:   end for
8:   if  $E|k$  then
9:     for  $i \in [N]$  do
10:      server  $i$  does quantization  $d_k^i = Q(w_k^i - s_{k-E}^i, \frac{s\eta_k}{2})$ ;
11:      server  $i$  updates  $s_k^i = d_k^i + s_{k-E}^i$ ;
12:      server  $i$  send  $d_k^i$  to server 0;
13:    end for
14:    server 0 computes  $s_k = \sum_{i=1}^N p_i d_k^i + s_{k-E}$ ;
15:    server 0 does quantization  $u_k = Q(s_k - v_{k-E}, \frac{s\eta_k}{2})$ ;
16:    server 0 sends  $u_k$  to server  $i, \forall i \in [N]$ ;
17:    server 0 updates  $v_k = u_k + v_{k-E}$ ;
18:    for  $i \in [N]$  do
19:      server  $i$  updates  $v_k = u_k + v_{k-E}$ ;
20:      server  $i$  updates  $w_k = v_k$ ;
21:    end for
22:  end if
23: end while

```

D.1 Proof for DEED-GD

Theorem D.1. In algorithm 1, we choose the learning rate $\eta_t \equiv \eta = \frac{2}{L+\mu}$, $c := 1 - \eta\mu$, $c < c' < 1$, and the maximal error at iteration t is $sc'^{t+1}/2$ where s is the quantization level. Then algorithm 1 communicates $\tilde{O}(Nd)$ bits at iteration $t \geq 1$, and

$$\|w_t - w^*\| \leq (c')^t \left(\max \left\{ 0, \|w_0 - w^*\| - \frac{c\eta s}{c' - c} \right\} + \frac{c'\eta s}{c' - c} \right). \quad (10)$$

Proof sketch. First of all, according to mathematical induction and triangle inequality, we see that the effective at iteration t is bounded by $\eta \cdot sc'^{t+1}$. Besides, consider the function $F_t(w) \equiv F(w) := w - \eta \nabla f(w)$. Because f is L -smooth and μ -convex, we know that F is $(1 - \eta\mu)$ -Lipschitz with fixed point w^* . Then, according to our framework, we can prove that $\|w_t - w^*\|$ converges at speed c' . This is exactly (10).

To bound the number of bits produced by each communication, we only need to prove that $\|g_t^i - s_{t-1}^i\| = \Theta(c'^t)$ and $\|s_t - v_{t-1}\| = \Theta(c'^t) \forall i \in [N]$ because of lemma 2.2.

Notice that these two norms are all close to $L\|w_t - w_{t-1}\|$, and $\|w_t - w_{t-1}\|$ can be bounded by $\|w_t - w^*\| + \|w^* - w_{t-1}\|$ which are also $\Theta(c'^t)$ by (10) and everything's done. In our real proof, we will bound these terms carefully to give a tighter bound.

Proof. First of all, we can give a deterministic version of theorem 2.4 as we promised below theorem 2.4.

Proposition D.2. Suppose $F : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a continuous function with Lipschitz constant $c < 1$. Define a sequence $\{e_i\}_{i \geq 1}$ in \mathbb{R}^d satisfies $\|e_i\| \leq \eta s c'^i$ where $c < c' < 1$. Then $\forall x_0 \in \mathbb{R}^d$, the sequence constructed by $x_{t+1} = F(x_t) + e_{t+1}$ satisfies $D(x_t) \leq c'^t \left(\max \left\{ 0, D(x_0) - \frac{c\eta s}{c'-c} \right\} + \frac{c'\eta s}{c'-c} \right)$ where $D(w) := \|w - w^*\|$ and w^* is the fixed point of F .

Proof. According to definition, we have the following inequalities.

$$\begin{aligned} D(x_{k+1}) &= D(F(x_k) + e_{k+1}) \\ &\leq D(F(x_k)) + \|e_{k+1}\| \\ &\leq c \cdot D(x_k) + \eta s c'^{k+1}. \end{aligned}$$

Hence,

$$\begin{aligned} \frac{D(x_k)}{c^k} &\leq \frac{D(x_{k-1})}{c^{k-1}} + \eta s \left(\frac{c'}{c} \right)^k \\ &\leq D(x_0) + \eta s \sum_{i=1}^k \left(\frac{c'}{c} \right)^i \\ &= D(x_0) + \eta s \frac{\left(\frac{c'}{c} \right)^{k+1} - 1}{\frac{c'}{c} - 1} \end{aligned}$$

Finally, we have $D(x_k) \leq c^k \left(D(x_0) - \frac{c\eta s}{c'-c} \right) + c'^k \left(\frac{c'\eta s}{c'-c} \right)$ which is bounded by $c'^k \left(\max \left\{ 0, D(x_0) - \frac{c\eta s}{c'-c} \right\} + \frac{c'\eta s}{c'-c} \right)$. \square

Then we can prove the inequality (10). Notice $\nabla F = I - \eta \nabla^2 f$ where $\mu I \preceq \nabla^2 f \preceq LI$, we see F is $c := 1 - \eta\mu$ Lipschitz. By proposition D.2, we only need to prove $\|v_k - g_k\| \leq s c'^{k+1}$ where $g_k = \nabla f(x_k)$. By induction, we have $s_k - \frac{1}{N} \sum_{i=1}^N s_k^i = \frac{1}{N} \sum_{i=1}^N d_k^i + s_{k-1} - \frac{1}{N} \sum_{i=1}^N (d_k^i + s_{k-1}^i) = s_{k-1} - \frac{1}{N} \sum_{i=1}^N s_{k-1}^i = s_{-1} - \frac{1}{N} \sum_{i=1}^N s_{-1}^i = 0$. Then,

$$\begin{aligned} \|v_k - g_k\| &= \|u_k + v_{k-1} - g_k\| \\ &\leq \|u_k - (s_k - v_{k-1})\| + \|s_k - g_k\| \\ &\leq \|u_k - (s_k - v_{k-1})\| + \left\| \frac{1}{N} \sum_{i=1}^N (d_k^i + s_{k-1}^i - g_k^i) \right\| \\ &\leq \|u_k - (s_k - v_{k-1})\| + \frac{1}{N} \sum_{i=1}^N \|d_k^i + s_{k-1}^i - g_k^i\| \\ &\leq s c'^{k+1}. \end{aligned}$$

For convenience, we define constant $\xi_s = \max \left\{ 0, \|x_0 - x^*\| - \frac{c\eta s}{c'-c} \right\} + \frac{c'\eta s}{c'-c}$. The convergence result comes directly from proposition D.2.

To bound the number of bits, we only need to calculate the maximal norm of the vector we need to encode. Actually, $\forall i \in [N], \forall k \geq 1$ we have

$$\|g_k^i - s_{k-1}^i\| \leq \|g_k^i - g_{k-1}^i\| + \|g_{k-1}^i - s_{k-1}^i\|$$

$$\begin{aligned}
&\leq L\|x_k - x_{k-1}\| + \frac{sc'^k}{2} \\
&\leq L\eta\|v_{k-1}\| + \frac{sc'^k}{2} \\
&\leq L\eta\|g_{k-1}\| + L\eta sc'^k + \frac{sc'^k}{2} \\
&\leq L^2\eta\|x_{k-1} - x^*\| + L\eta sc'^k + \frac{sc'^k}{2} \\
&\leq L^2\eta c'^{k-1}\xi_s + L\eta sc'^k + \frac{sc'^k}{2} \\
&= \left(\frac{L^2\eta\xi_s + L\eta sc' + \frac{sc'}{2}}{c'^2} \right) c'^{k+1}.
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\|s_k - v_{k-1}\| &\leq \|s_k - g_k\| + \|g_k - g_{k-1}\| + \|g_{k-1} - v_{k-1}\| \\
&\leq L\|x_k - x_{k-1}\| + \frac{3sc'^k}{2} \\
&\leq \left(\frac{L^2\eta\xi_s + L\eta sc' + \frac{3sc'}{2}}{c'^2} \right) c'^{k+1}.
\end{aligned}$$

In this case, the error fraction (length divided by error, i.e. inverse of relative error) is bounded by $\zeta_s := \frac{2L^2\eta\xi_s + 2L\eta c' + 3c'}{c'^2}$, and the number of bits is at most $(1.05 + \log_2(\zeta_s + 2))d$ by lemma 2.2. \square

D.2 Proof for DEED-SGD

Theorem D.3. *In algorithm 2, we assume (Weak Growth Condition) WGC is satisfied for approximate gradient $\bar{\nabla}_i$ for every f_i with parameter ρ . We choose the learning rate $\eta_t \equiv \eta = \frac{1}{\rho L}$, $c := 1 - \eta\mu$, $c < c' < 1$, and the maximal error at iteration t is $\sqrt{sc'^{k+1}}/2$ and the error is unbiased.*

DEED-SGD communicates $\tilde{O}(Nd)$ bits at iteration $t \geq 1$, and

$$\mathbb{E}\|w_t - w^*\|^2 \leq (c')^t \left(\max \left\{ 0, \|w_0 - w^*\|^2 - \frac{c\eta^2 s}{c' - c} \right\} + \frac{c'\eta^2 s}{c' - c} \right). \quad (11)$$

Proof sketch. First of all, we introduce theorem 5 in [21] to show that with WGC, stochastic gradient descent converges linearly.

Lemma D.4. *Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is μ -strongly-convex. Besides, $f(x) = \frac{1}{N} \sum_{i=1}^N f_i(x)$, where each f_i is L -smooth. We assume WGC is satisfied for approximate gradient $\bar{\nabla}_i$ with parameter ρ . Then series $\{x_i\}_{i \geq 0}$ generated by iteration formula*

$$x_{k+1} := x_k - \frac{1}{N} \sum_{i=1}^N \bar{\nabla}_i f_i(x_k) \quad (12)$$

satisfy

$$\mathbb{E}[\|x_{k+1} - x^*\|^2 | x_k] \leq \left(1 - \frac{\mu}{\rho L} \right) \|x_k - x^*\|^2.$$

Being similar to the proof in previous section, we can bound the effective error by $\eta\sqrt{sc'^{k+1}}$. Then we have

$$\mathbb{E}\|x_{k+1} - x_k\|^2 \leq c\mathbb{E}\|x_k - x^*\|^2 + \eta^2 sc'^{k+1}.$$

Then we can put it in our framework. By using the same technique in proposition D.2, we have inequality (11). Finally, we can bound the number of bits by using lemma 2.2.

Proof. As we showed in sketch, we have proved inequality (11). We only need to prove that $\mathbb{E}\|g_k^i - s_{k-1}^i\|^2$ and $\mathbb{E}\|s_k - v_{k-1}\|^2$ are $O(c'^k)$, $\forall i \in [N]$, because $\ln x^2 = 2 \ln x$ is a concave function. By triangle inequality and the definition of L -smooth, we only need to prove $\mathbb{E}\|w_k - w_{k-1}\|^2 \leq 2\mathbb{E}\|w_k - w^*\|^2 + 2\mathbb{E}\|w^* - w_{k-1}\|^2$ is $O(c'^k)$, which is obvious by inequality (11). \square

D.3 Proof for DEED-Fed

Theorem D.5. In algorithm 3, we choose the learning rate $\eta_t := \frac{\beta}{t+\gamma}$ for some $\beta > \frac{1}{\mu}$, $\gamma > 1$ such that $\eta_0 \leq \frac{1}{4L}$ and $\eta_t \leq 2\eta_{t+E}$. Let the maximal error at iteration $t \in \{0, E, 2E, \dots\}$ be $s\eta_t$. Then DEED-Fed communicates $\tilde{O}(Nd)$ bits at iteration $t \in \{E, 2E, \dots\}$, and

$$\mathbb{E}\|\bar{w}_t - w^*\|^2 \leq \frac{v}{\gamma + t}, \quad (13)$$

where $v := \max \left\{ \frac{\beta^2(B+C+s^2)}{\beta\mu-1}, \gamma\|w_0 - w^*\|^2 \right\}$. Here, $B = \sum_{k=1}^N p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2$, and C is a constant depending on different setting. In full participant setting, $C = 0$. In partial participant setting 1, $C = \frac{4}{K} E^2 G^2$, and in partial participant setting 1, $C = \frac{N-K}{N-1} \frac{4}{K} E^2 G^2$.

Proof sketch. Recall the theorem with no quantization in [41].

Theorem D.6. Assume assumption C.1 and C.2 hold. For FedAvg, we have

$$\mathbb{E}\|\bar{w}_{t+1} - w^*\|^2 \leq (1 - \eta_t \mu) \mathbb{E}\|\bar{w}_t - w^*\|^2 + \eta_t^2 (B + C), \quad (14)$$

where $B = \sum_{k=1}^N p_k^2 \sigma_k^2 + 6L\Gamma + 8(E-1)^2 G^2$, and C is a constant depending on different setting. In full participant setting, $C = 0$. In partial participant setting 1, $C = \frac{4}{K} E^2 G^2$, and in partial participant setting 1, $C = \frac{N-K}{N-1} \frac{4}{K} E^2 G^2$.

Moreover, with inequality (14), we have $\mathbb{E}\|\bar{w}_t - w^*\|^2 \leq \frac{v}{\gamma+t}$ where $v := \max \left\{ \frac{\beta^2(B+C)}{\beta\mu-1}, \gamma\|w_0 - w^*\|^2 \right\}$.

Combine this theorem error analysis, we have⁵

$$\mathbb{E}\|\bar{w}_{t+1} - w^*\|^2 \leq (1 - \eta_t \mu) \mathbb{E}\|\bar{w}_t - w^*\|^2 + \eta_t^2 (B + C + s^2).$$

We can already put the map from \bar{w}_t to \bar{w}_{t+1} into our framework, and prove that it converges sublinearly. Actually, we can just use theorem 14 and conclude that inequality (13) holds. The only difficulty is the bound for communication. We cannot bound $\|\bar{w}_t - \bar{w}_{t+E}\|$ by $\|\bar{w}_t - w^*\| + \|w^* - \bar{w}_{t+E}\|$ since it is $O(1/\sqrt{t})$, while the precision is $O(1/t)$. Please see this part in the proof below.

⁵We can prove better inequality since the quantization is only done on iteration $E, 2E, \dots$. However, this is enough since it won't change the convergence rate.

Proof. We have proved inequality (13). The only thing left is to prove that $\mathbb{E}\|w_{(t+1)E}^k - w_{tE}^k\|^2 = O(1/t^2)$.

Notice that for each larger iteration $t \geq 0$ and slave node k , we have

$$\begin{aligned} \|w_{(t+1)E}^k - w_{tE}^k\| &\leq \sum_{i=0}^{E-1} \|w_{tE+i+1}^k - w_{tE+i}^k\| \\ &\leq \sum_{i=0}^{E-1} \eta_{tE+i} \|\nabla F_k(w_{tE+i}^k, \xi_{tE+i}^k)\| \\ &\leq \eta_{tE} \sum_{i=0}^{E-1} \|\nabla F_k(w_{tE+i}^k, \xi_{tE+i}^k)\|. \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{E}\|w_{(t+1)E}^k - w_{tE}^k\|^2 &\leq \eta_{tE}^2 \cdot E \sum_{i=0}^{E-1} (\mathbb{E}\|\nabla F_k(w_{tE+i}^k)\|^2 + \sigma_k^2) \\ &\leq \eta_{tE}^2 \cdot E \sum_{i=0}^{E-1} (L^2 \mathbb{E}\|w_{tE+i}^k - w^*\|^2 + \sigma_k^2) \\ &\leq \eta_{tE}^2 \cdot E \sum_{i=0}^{E-1} (2L^2 \mathbb{E}\|w_{tE+i}^k - \bar{w}_{tE+i}\|^2 + 2L^2 \mathbb{E}\|\bar{w}_{tE+i} - w^*\|^2 + \sigma_k^2) \\ &\leq \eta_{tE}^2 \cdot E \sum_{i=0}^{E-1} \left(2L^2 \eta_{tE}^2 E^2 G^2 + 2L^2 \frac{v}{\gamma + tE} + \sigma_k^2 \right) \\ &= \eta_{tE}^2 \cdot E^2 \left(2L^2 \eta_{tE}^2 E^2 G^2 + 2L^2 \frac{v}{\gamma + tE} + \sigma_k^2 \right). \end{aligned}$$

With the approximation above, all the vectors we need to do quantization are bounded by

$$\sqrt{\eta_{tE}^2 s^2 + \eta_{tE}^2 \cdot E^2 \left(2L^2 \eta_{tE}^2 E^2 G^2 + 2L^2 \frac{v}{\gamma + tE} + \sigma_k^2 \right)}$$

in expectation, and proof follows. \square

E Algorithm A-DEED-GD and its Convergence Analysis

A-DEED-GD is the accelerated version in DEED series. The algorithm and its proof are similar to DEED-GD. The only difference between DEED-GD is the update rule. Please see algorithm 4 below for details.

Proof sketch First of all, we can use triangle inequality to prove that the error on v_k is small, i.e. $\|g_k - v_k\| \leq sc'^{k+1}$. Then, with diminishing error we are able to prove linear convergence. Finally, we use lemma 2.2 to show that we communicate $O(d)$ bits per communication. The first and the third step are exactly the same as DEED-GD. Hence, we only need to prove that the convergence part.

Proof. Suppose $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a μ -convex and L -smooth function.

Choose an arbitrary point $x_0 = y_0 = v_0 \in \mathbb{R}^d$, we can define $\phi_0(x) := \phi_0^* + \frac{\mu}{2} \|x - v_0\|^2$ where $\phi_0^* := f(v_0)$. Then by definition we know $\phi(x) \leq f(x)$ and $\phi_0^* \geq f(x_0)$.

Algorithm 4: A-DEED-GD

```

1: Initialization: Each server  $i \in [N]$  holds  $x_0 = s_{-1}^i = v_{-1} = 0$ , server 0 holds  $v_{-1} = 0$   $k = 0$ ;
2: Parameter setting:  $\eta = \frac{1}{L}$ ,  $\tau = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ ,  $c = \sqrt{1 - \sqrt{\frac{\mu}{L}}}$ ;
3: while the precision is not enough do
4:   for  $i \in [N]$  do
5:     server  $i$  computes  $g_k^i = \nabla f_i(y_k)$ ;
6:     server  $i$  does quantization  $d_k^i = Q(g_k^i - s_{k-1}^i, \frac{sc'^{k+1}}{2})$ ;
7:     server  $i$  updates  $s_k^i = d_k^i + s_{k-1}^i$ ;
8:     server  $i$  send  $d_k^i$  to server 0;
9:   end for
10:  server 0 computes  $s_k = \frac{1}{N} \sum_{i=1}^N d_k^i + s_{k-1}$ ;
11:  server 0 does quantization  $u_k = Q(s_k - v_{k-1}, \frac{sc'^{k+1}}{2})$ ;
12:  server 0 sends  $u_k$  to server  $i, \forall i \in [N]$ ;
13:  server 0 updates  $v_k = u_k + v_{k-1}$ ;
14:  for  $i \in [N]$  do
15:    server  $i$  updates  $v_k = u_k + v_{k-1}$ ;
16:    server  $i$  updates  $x_{k+1} = y_k - \eta v_k$ ;
17:    server  $i$  updates  $y_{k+1} = x_{k+1} + \tau(x_{k+1} - x_k)$ ;
18:  end for
19:   $k = k + 1$ ;
20: end while

```

Next, we inductively define the following quantity. Suppose $\ell : \mathbb{Z}^{\geq 0} \rightarrow \mathbb{R}^+$ is an arbitrary function.

$$\begin{aligned}
x_{k+1} &= y_k - \frac{1}{L} m_k \\
\phi_{k+1}(x) &= (1 - \alpha) \phi_k(x) + \alpha \left[f(y_k) + \langle m_k, x - y_k \rangle + \frac{\mu}{2} \|x - y_k\|^2 \right] \\
v_{k+1} &= \arg \min_{v \in \mathbb{R}^n} \phi_{k+1}(v) \\
y_{k+1} &= \frac{x_{k+1} + \alpha v_{k+1}}{1 + \alpha} \\
\phi_{k+1}^* &= \min \phi_{k+1}
\end{aligned}$$

where $\alpha = \sqrt{\frac{\mu}{L}}$ and $m_k \in \mathbb{R}^d$ such that $\|m_k - \nabla f(y_k)\| \leq c^{k+1} \ell(k)$.

Besides, we will construct monotonically increasing functions $h, g : \mathbb{Z}^{\geq 0} \rightarrow \mathbb{R}$ inductively such that

$$\begin{aligned}
f(x_k) &\leq \phi_k^* + h(k) \cdot c^{2k} \\
\phi_k(x^*) &\leq (1 - c^{2k}) f^* + c^{2k} \phi_0(x^*) + g(k) \cdot c^{2k}.
\end{aligned}$$

Obviously, it is appropriate to set $g(0) = h(0) = 0$.

Before we go deeper, here is an important lemma.

Lemma E.1. *With the definition above, we have*

$$y_{k+1} = x_{k+1} + \tau(x_{k+1} - x_k)$$

where $\tau = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$.

Proof. Because $\phi_k(x) = \phi_k^* + \frac{\mu}{2}\|x - v_k\|^2$, taking derivative $\nabla\phi_{k+1}(x) = \mu(1-\alpha)(x - v_k) + \alpha m_k + \alpha\mu(x - y_k)$. And then we get $v_{k+1} = (1-\alpha)v_k + \alpha y_k - \frac{\alpha}{\mu}m_k$.

Because $y_k = \frac{x_k + \alpha v_k}{1+\alpha}$, we can substitute $v_k = \frac{1+\alpha}{\alpha}y_k - \frac{1}{\alpha}x_k$ and get

$$\begin{aligned} v_{k+1} &= \frac{1-\alpha^2}{\alpha}y_k - \frac{1-\alpha}{\alpha}x_k + \alpha y_k - \frac{\alpha}{\mu}m_k \\ &= \frac{1}{\alpha}\left(y_k - \frac{1}{L}m_k\right) - \frac{1-\alpha}{\alpha}x_k \\ &= \frac{x_{k+1} - x_k}{\alpha} + x_k. \end{aligned}$$

Hence

$$\begin{aligned} y_{k+1} &= \frac{x_{k+1} + \alpha v_{k+1}}{1+\alpha} \\ &= x_{k+1} + \tau(x_{k+1} - x_k). \end{aligned}$$

□

Then we have

$$\begin{aligned} f(x_k) - f^* &\leq \phi_k^* + h(k) \cdot c^{2k} - f^* \\ &\leq c^{2k}(\phi_0(x^*) - f^* + g(k) + h(k)) \\ &= c^{2k}(\Delta + g(k) + h(k)) \end{aligned}$$

where $\Delta := \phi_0(x^*) - f^* = f(v_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2$. Hence,

$$\begin{aligned} \|x_k - x^*\| &\leq \sqrt{\frac{2}{\mu}(f(x_k) - f^*)} \\ &= \sqrt{\frac{2}{\mu}} \cdot c^k \cdot \sqrt{\Delta + g(k) + h(k)}. \end{aligned} \tag{15}$$

Furthermore, according to lemma 3 and the monotony of g and h , we have

$$\begin{aligned} \|y_k - x^*\| &= \left\| \frac{2\sqrt{L}}{\sqrt{L} + \sqrt{\mu}}(x_k - x^*) - \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}(x_{k-1} - x^*) \right\| \\ &\leq \left(\frac{2\sqrt{L}}{\sqrt{L} + \sqrt{\mu}} + \frac{1}{c} \frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right) \sqrt{\frac{2}{\mu}} \cdot c^k \cdot \sqrt{\Delta + g(k) + h(k)} \\ &\leq 3\sqrt{\frac{2}{\mu}} \cdot c^k \cdot \sqrt{\Delta + g(k) + h(k)}. \end{aligned}$$

And then we can give the first upper bound

$$\phi_{k+1}(x^*) = (1-\alpha)\phi_k(x^*) + \alpha \left[f(y_k) + \langle m_k, x^* - y_k \rangle + \frac{\mu}{2} \|x^* - y_k\|^2 \right]$$

$$\begin{aligned}
&\leq (1-\alpha)\phi_k(x^*) + \alpha \left[f(y_k) + \langle \nabla f(y_k), x^* - y_k \rangle + \frac{\mu}{2} \|x^* - y_k\|^2 \right] \\
&\quad + \alpha \cdot \ell(k) c^{k+1} \|x^* - y_k\| \\
&\leq (1-\alpha)\phi_k(x^*) + \alpha f^* + \alpha \cdot \ell(k) c^{k+1} \|x^* - y_k\| \\
&\leq c^2 \left((1-c^{2k})f^* + c^{2k}\phi_0(x^*) + g(k) \cdot c^{2k} \right) + (1-c^2)f^* \\
&\quad + \alpha \cdot \ell(k) c^{k+1} \|x^* - y_k\| \\
&= (1-c^{2k+2})f^* + c^{2k+2}\phi_0(x^*) + g(k) \cdot c^{2k+2} \\
&\quad + 3\ell(k) \sqrt{\frac{2}{L}} \cdot c^{2k+1} \cdot \sqrt{\Delta + g(k) + h(k)},
\end{aligned}$$

which means we only need to make

$$g(k+1) \geq g(k) + \frac{3\ell(k) \sqrt{\frac{2}{L}} \cdot \sqrt{\Delta + g(k) + h(k)}}{c} \quad (16)$$

To make an upper bound of $h(k+1)$, we notice that

$$\begin{aligned}
\phi_{k+1}^* &= \phi_{k+1}(v_{k+1}) \\
&= (1-\alpha) \left(\phi_k^* + \frac{\mu}{2} \|v_{k+1} - v_k\|^2 \right) + \alpha f(y_k) + \alpha \langle m_k, v_{k+1} - y_k \rangle \\
&\quad + \frac{\alpha\mu}{2} \|v_{k+1} - y_k\|^2.
\end{aligned}$$

Substitute $v_{k+1} - y_k = (1-\alpha)(v_k - y_k) - \frac{\alpha}{\mu} m_k$, we have

$$\begin{aligned}
\phi_{k+1}^* &= (1-\alpha)\phi_k^* + \alpha f(y_k) - \frac{1}{2L} \|m_k\|^2 \\
&\quad + \alpha(1-\alpha) \left(\frac{\mu}{2} \|y_k - v_k\|^2 + \langle m_k, v_k - y_k \rangle \right) \\
&\geq (1-\alpha) \left(f(x_k) - h(k) \cdot c^{2k} \right) + \alpha f(y_k) - \frac{1}{2L} \|m_k\|^2 \\
&\quad + \alpha(1-\alpha) \left(\frac{\mu}{2} \|y_k - v_k\|^2 + \langle m_k, v_k - y_k \rangle \right).
\end{aligned}$$

Because $f(x_k) \geq f(y_k) + \langle \nabla f(y_k), x_k - y_k \rangle \geq f(y_k) + \langle m_k, x_k - y_k \rangle - \ell(k) \cdot c^{k+1} \|x_k - y_k\|$ and $\|x_k - y_k\| = \tau \|x_k - x_{k-1}\| \leq c^2 (\|x_k - x^*\| + \|x^* - x_{k-1}\|) \leq c^{k+1} \cdot (1+c) \sqrt{\frac{2}{\mu}} \sqrt{\Delta + g(k) + h(k)}$.

Hence, we can bound ϕ_{k+1}^* by

$$\begin{aligned}
&f(y_k) - \frac{1}{2L} \|m_k\|^2 - h(k) \cdot c^{2k+2} \\
&- c^{2k+4} \cdot \ell(k) \cdot (1+c) \sqrt{\frac{2}{\mu}} \sqrt{\Delta + g(k) + h(k)}.
\end{aligned}$$

Recall that

$$\begin{aligned}
f(x_{k+1}) - f(y_k) + \frac{1}{2L} \|m_k\|^2 &\leq \frac{1}{L} \langle m_k - \nabla f(y_k), m_k \rangle \\
&\leq \frac{1}{L} \ell(k) \cdot c^{k+1} \|m_k\| \\
&\leq c^{2k+1} \cdot \left(\frac{c\ell^2(k)}{L} + 3\sqrt{\frac{2}{\mu}} \ell(k) \sqrt{\Delta + g(k) + h(k)} \right)
\end{aligned}$$

we have

$$\begin{aligned}\phi_{k+1}^* &\geq f(x_{k+1}) - c^{2k+1} \cdot \left(\frac{c\ell^2(k)}{L} + 3\sqrt{\frac{2}{\mu}}\ell(k)\sqrt{\Delta + g(k) + h(k)} \right) \\ &\quad - h(k) \cdot c^{2k+2} - c^{2k+4} \cdot \ell(k) \cdot (1+c)\sqrt{\frac{2}{\mu}}\sqrt{\Delta + g(k) + h(k)},\end{aligned}$$

which means $h(\cdot)$ only need to satisfy

$$h(k+1) \geq h(k) + \frac{5}{c}\ell(k) \cdot \sqrt{\frac{2}{\mu}}\sqrt{\Delta + g(k) + h(k)} + \frac{\ell^2(k)}{L}. \quad (17)$$

Finally, we only need to make inequalities (16) and (17) be the iteration formulas for arrays $\{g(k)\}_{k \geq 0}, \{h(k)\}_{k \geq 0}$ with initialization $g(0) = h(0) = 0$, and we have

$$i(k+1) \geq i(k) + \left(\frac{3\sqrt{\frac{2}{L}} + 5\sqrt{\frac{2}{\mu}}}{c} \right) \ell(k)\sqrt{\Delta + i(k)} + \frac{\ell^2(k)}{L} \quad (18)$$

where $i(k) := g(k) + h(k), \forall k \geq 0$. Now, $\forall c < c' < 1$, define $\gamma = \frac{c'}{c}$ and $\ell(k) = s\gamma^{k+1}$. We will prove that $i(k) \leq C\gamma^{2k}, \forall k \geq 0$ for sufficient large C by mathematical induction.

Obviously, it holds for index 0. Suppose it holds for index less or equal to k . Then we only need to prove

$$C\gamma^{2k+2} \geq \left(C + \frac{s^2}{L} \right) \gamma^{2k} + \left(\frac{3\sqrt{\frac{2}{L}} + 5\sqrt{\frac{2}{\mu}}}{c} \right) s\gamma^{k+1}\sqrt{\Delta + C\gamma^{2k}}.$$

This can be derived by

$$C(\gamma^2 - 1) \geq \frac{s^2}{L} + \left(\frac{3\sqrt{\frac{2}{L}} + 5\sqrt{\frac{2}{\mu}}}{c} \right) s\gamma\sqrt{\Delta + C}.$$

Hence, if we define $\alpha_s = \frac{s^2}{L(\gamma^2-1)} + \Delta$, $\beta_s = \left(\frac{3\sqrt{\frac{2}{L}} + 5\sqrt{\frac{2}{\mu}}}{c(\gamma^2-1)} \right) s\gamma$, making $C \geq \beta_s^2 + \alpha_s + \beta_s\sqrt{\alpha_s} - \Delta$ is enough.

The convergence result follows by inequality (15). \square

F Experiments

F.1 Linear regression

We illustrate the effectiveness of our proposed algorithms (DEED-GD and its accelerated version A-DEED-GD) in frequent-communication large-memory setting on linear regression problem with a 100 by 100 Gaussian generated matrix with condition number equals to 16. We focus on star networks, and there are 10 computing nodes. We perform 800 epochs on non-accelerated algorithms and 200 epochs on accelerated ones. In each update, the stepsize chosen on each computing node i is $\min_i 1/L_i$ where L_i is the Lipschitz constant for the function corresponds to node i . As required in theorem 3.4 in QSGD's paper [13], we choose another learning rate for QSGD. The experiments are done on a computer with 2 GHz Dual-Core Intel Core i5 processor.

We choose the quantization level to be 10000 in QSGD [13] as smaller quantization level lead to larger loss values. For DIANA, as suggested by [19], we quantize the weight of each layer separately, and we use either the full block size of the weights or let block size equal to 20. The quantization level in DIANA algorithms is equal to the block size d_i for some vector i and the parameter α in DIANA is chosen to be $\min_i 1/\sqrt{d_i}$. For DEED-GD, we choose $s = 0.01$ and $c = 0.95$ which is the convergence parameter of the baseline method (GD) using a stepsize $\min_i 1/L_i$. Here s, c are parameter controlling the maximal error defined below. For A-DEED-GD, we choose $s = 0.1$ and $c = 0.82$ which is the convergence parameter of the accelerated baseline method (A-GD) using a stepsize $\min_i 1/L_i$.

In our algorithms, to do quantization with maximal error $\varepsilon (= s \cdot c^{k+1})$ at iteration k , we consider the following algorithm. For a vector w , we first compute a vector $\tilde{w} := \frac{w\sqrt{d}}{\varepsilon}$. Next, we encode each coordinate of \tilde{w} , say \tilde{w}_i into $\lfloor \tilde{w}_i \rfloor$ and $\lfloor \tilde{w}_i \rfloor + 1$ unbiasedly. We call the new vector \tilde{v} . Finally, we compute the quantized vector $v := \frac{\tilde{v}\varepsilon}{\sqrt{d}}$. It is obvious that using this quantization method the error is bounded by ε .

To encode the integer vector \tilde{v} into an integer, we use the method introduced in QSGD. We refer Elias encoding, a map from positive integer to non-negative integer. First of all, We use Elias to encode the first non-zero element of \tilde{v} (use 1 bit to encode sign and use other bits to encode absolute value) and its position. Then we iteratively use Elias to encode the next non-zero element and its distance from the previous position. It works perfectly especially when \tilde{v} is sparse.

As mentioned in table 1, we noticed that there could be two settings for algorithm only does quantization on computing nodes like DIANA. In star network setting, the center node transmits full vector (with no quantization) to computing nodes in broadcasting term. This would cost extra $32dN$ bits. In fully connected network setting, computing nodes broadcast to each other and update information separately. This would lead to an extra $N - 1$ factor on total number of bits since each computing node should broadcast to $N - 1$ other nodes. One another method is to time 2 instead of time $N - 1$ on total number of bits from computing nodes to center node to compare them with DEED series. It is reasonable for the following reasons. 1) In DEED series, we have two communications for each computing nodes in each iteration: sending message and receiving message. So, in this case what we really compare is proportional to number of iterations \times number of bits per communication. 2) The “ $\times 2$ ” scheme counts less bits than both star network setting and fully connected network setting. If we can beat DIANA and other algorithms in this setting, it means our framework is essentially better than their framework, i.e. better than them in any settings. Notice that this is only a method of counting bits, not a method of communication.

The performance analysis has already provided in Section 5 and it shows that our algorithms (DEED-GD and its accelerated version A-DEED-GD) save the most number of bits in communication without scarifying the convergence speed. We also run the experiments on 5 different random seeds and the results is shown in Figure 2. In Figure 2, the shaded regions line up with the maximal and minimal loss values at each epoch among the 5 different runs, and the regions are too small to visualize due to small variance.

F.2 Image classification on MNIST

We evaluate the effectiveness of the proposed algorithms via training a neural network on the MNIST dataset [42] for image classification. MNIST consists of 60,000 28×28 pixel training images containing a single numerical digit and an additional 10,000 test examples. Our neural network consists of one 500-neuron fully-connected layer followed by a ten unit softmax layer for classification, and the layer used ReLU activations [44]. The experiments are performed on a using NVIDIA GeForce GTX1080 GPU, and the models are distributed over 6 computing servers, where each of the servers have access to 10,000 training images. In large-memory setting, each server uses its own 10,000 images to update the models, while in small-memory setting, each

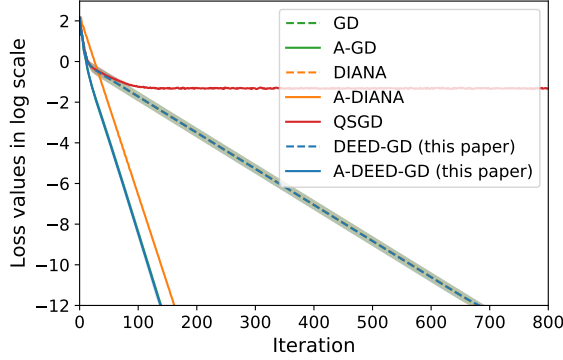


Figure 2: Compare loss values of each algorithms among 5 different runs

server uses randomly selected 1666 images among its own 10,000 training images as a minibatch. We train the models for 200 epochs in large-memory setting and 100 epochs for small-memory setting. The training time is approximately 9 hours for each algorithm in small-memory setting and 6 hours for large-memory setting.

We compare our algorithms with QSGD [13], DIANA [19], DoubleSqueeze [25], and Terngrad [14] in frequent communication settings. For all the algorithms except DIANA, in every update, we vectorize the gradient or the gradient difference of the weight matrices of the neural network, concatenate these vectors as a large vector and do quantization, and reshape the quantized vector into the original shapes of the weight matrices for updating. For DIANA, as suggested by [19], we quantize the weight of each layer separately, and we use two different block sizes in quantization (use the full block size of the weights, or let block size equal to 128). As suggested by [19], the quantization level in DIANA algorithms is equal to the block size d_i for some vector i and the parameter α in DIANA is chosen to be $\min_i 1/\sqrt{d_i}$. In large-memory setting, we train our algorithm DEED-GD with the parameters $e = 0.1, s = 25$, and in small-memory setting, we train DEED-GD with parameters $e = 0.2, s = 25$. Here the maximal error $\varepsilon := \frac{s}{(k+1)^e}$. We use the same encoding algorithm as described in linear regression. In both large-memory and small-memory setting, we present the results for choosing 4-bit quantization for QSGD. For DoubleSqueeze, we perform two kinds of quantization as discussed in [25]: top-k compression and 1-bit compression. For fair comparison, we use the same stepsize for all the algorithms, where the stepsize for large-memory setting is 0.25 and is 1.18 for small-memory setting. These stepsizes are chosen as the loss curves of the baseline methods (GD and SGD) are smooth and the baseline methods achieve fairly high accuracy in testing. As shown in Figure 3(a) and Figure 3(b), our algorithms as well as QSGD, DIANA and Terngrad achieve the same loss values as the baseline methods (GD and SGD) in both large-memory setting and small-memory setting.

To compare the efficiency of each algorithm, we compute the total number of bits throughout the training. As large integers are less frequent in encoded vectors [13], we use Elias integer encoding to save bits in communication [45] for all algorithms in comparison. Notice that Terngrad, DIANA and QSGD only perform quantization on computing nodes and will typically use 32-bit precision to encode the vectors which are sent from the center node in a star network. To be fair in comparison, we use the scaling technique proposed in [14] and use $\log_2(1 + 2 * N) * d$ bits to encode the vectors sent from center node where N is number of computing servers and d is the dimension of the vector. This number is significantly smaller than $32 * d$ unless $N \geq 2^{30}$. For DIANA and QSGD, we let computing nodes to share information to each other so they do not need to broadcast via the center node which will cause $32 * d$ bits for each update. Under this setting, the bits communicated in each update is $B * (N - 1)$ where B is the number of bits to communicated from

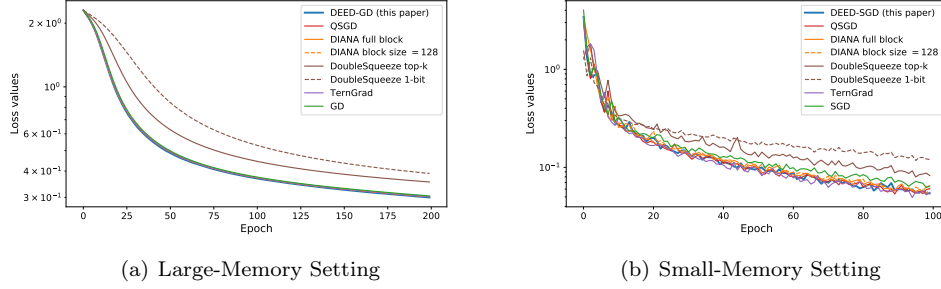


Figure 3: Compare loss values of each algorithms

the computing node and $N - 1$ is the number of other computing nodes that need to communicate to. In most cases, the bits computed in this way is fewer than using 32-bit precision to encode the vectors which are broadcast from the center node.

Results of Efficiency. To illustrate the efficiency of our proposed algorithms, we plot the number of bits vs the testing accuracy in Figure 4(a) and 4(b) for frequent large-memory setting and frequent small-memory setting. In both figures, the curve that is located on the left-most corresponds to our proposed algorithms, DEED-GD and DEED-SGD. This means that our proposed algorithms use the fewest total number of bits to achieve the accuracy.

To better illustrate how many bits we can save from other algorithms, we present the total number of bits in Table 3 and Table 4 and compute the ratio between the number of bits required by other algorithms and the number of bits required by the proposed algorithms in the fourth column of the Table. For example, in Table 4(a), the number 10.44 means DEED-GD takes 10.44 times fewer the number of bits than QSGD to achieve the accuracy. In theory, DIANA and our proposed algorithm DEED-GD both have linear convergence rate, but our experiments show that we can take 190.28 times fewer the number of bits than DIANA (block size equals to 128) to achieve the similar performance in training. In addition, our algorithms achieve the highest testing accuracy in the final epoch as shown in the second column of the table, and the accuracy is comparable or even higher than the ones achieved by the non-quantized baseline algorithms (GD achieves 91.7% and SGD achieves 97.37%).

To ensure the performance are reproducible, we also train the models under different random seeds and choose different parameters within certain ranges. For example, we train the models by our proposed algorithm using the parameters e is chosen from $[0.1, 0.3]$, s is chosen from $\{16, 25, 32\}$. The comparison we discussed above is still valid under these changes. We also run experiments for different quantization levels (e.g. using 2-bit quantization or 3-bit quantization) for QSGD and DIANA, but they cannot achieve the same testing accuracy with the same number of epochs as using 4-bit quantization, so we do not discuss these results here.

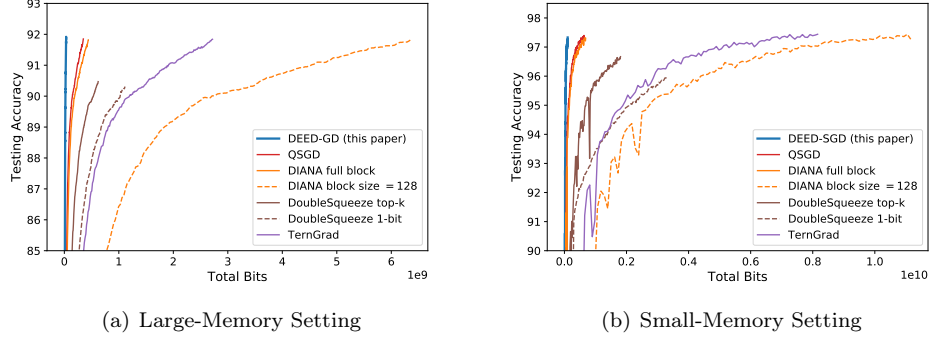


Figure 4: Total number of bits to reach certain testing accuracy

Algorithm	Testing accuracy	Total number of bits	Ratio
DEED-GD	91.86%	3.34×10^7	1.00
QSGD [13]	91.85%	3.49×10^8	10.44
DIANA (full block) [19]	91.82%	4.40×10^8	13.17
DIANA (block size = 128) [19]	91.82%	6.35×10^9	190.28
DoubleSqueeze top-k [25]	90.47%	6.21×10^8	18.59
DoubleSqueeze 1-bit [25]	90.31%	1.11×10^9	33.34
TernGrad [14]	91.84%	2.72×10^9	81.44

Table 3: DEED-GD saves bits in communication for large-memory setting

Algorithm	Testing accuracy	Total number of bits	Ratio
DEED-SGD	97.33%	1.04×10^8	1.00
QSGD [13]	97.31%	6.52×10^8	6.25
DIANA (full block) [19]	97.30%	6.83×10^8	6.55
DIANA (block size = 128) [19]	97.27%	1.12×10^{10}	106.95
DoubleSqueeze top-k [25]	96.67%	1.81×10^9	17.31
DoubleSqueeze 1-bit [25]	95.96%	3.34×10^9	32.01
TernGrad [14]	97.44%	8.16×10^9	78.20

Table 4: DEED-SGD saves bits in communication for small-memory setting

References

- [1] D. P. Bertsekas and J. N. Tsitsiklis, “Gradient convergence in gradient methods with errors,” *SIAM Journal on Optimization*, vol. 10, no. 3, pp. 627–642, 2000.
- [2] B. Recht, C. Re, S. Wright, and F. Niu, “Hogwild: A lock-free approach to parallelizing stochastic gradient descent,” in *Advances in neural information processing systems*, pp. 693–701, 2011.
- [3] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, *et al.*, “Large scale distributed deep networks,” in *Advances in neural information processing systems*, pp. 1223–1231, 2012.
- [4] A. Coates, B. Huval, T. Wang, D. Wu, B. Catanzaro, and N. Andrew, “Deep learning with cots hpc systems,” in *International conference on machine learning*, pp. 1337–1345, 2013.
- [5] M. Li, D. G. Andersen, J. W. Park, A. J. Smola, A. Ahmed, V. Josifovski, J. Long, E. J. Shekita, and B.-Y. Su, “Scaling distributed machine learning with the parameter server,” in *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 583–598, 2014.
- [6] T. Chilimbi, Y. Suzue, J. Apacible, and K. Kalyanaraman, “Project adam: Building an efficient and scalable deep learning training system,” in *11th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 14)*, pp. 571–582, 2014.
- [7] E. P. Xing, Q. Ho, W. Dai, J. K. Kim, J. Wei, S. Lee, X. Zheng, P. Xie, A. Kumar, and Y. Yu, “Petuum: A new platform for distributed machine learning on big data,” *IEEE Transactions on Big Data*, vol. 1, no. 2, pp. 49–67, 2015.
- [8] T. Chen, M. Li, Y. Li, M. Lin, N. Wang, M. Wang, T. Xiao, B. Xu, C. Zhang, and Z. Zhang, “Mxnet: A flexible and efficient machine learning library for heterogeneous distributed systems,” *arXiv preprint arXiv:1512.01274*, 2015.
- [9] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” *arXiv preprint arXiv:1603.04467*, 2016.
- [10] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally, “Deep gradient compression: Reducing the communication bandwidth for distributed training,” *arXiv preprint arXiv:1712.01887*, 2017.
- [11] F. Seide, H. Fu, J. Droppo, G. Li, and D. Yu, “1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [12] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan, “Deep learning with limited numerical precision,” in *International Conference on Machine Learning*, pp. 1737–1746, 2015.
- [13] D. Alistarh, D. Grubic, J. Li, R. Tomioka, and M. Vojnovic, “Qsgd: Communication-efficient sgd via gradient quantization and encoding,” in *Advances in Neural Information Processing Systems*, pp. 1709–1720, 2017.
- [14] W. Wen, C. Xu, F. Yan, C. Wu, Y. Wang, Y. Chen, and H. Li, “Terngrad: Ternary gradients to reduce communication in distributed deep learning,” in *Advances in neural information processing systems*, pp. 1509–1519, 2017.

- [15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, *et al.*, “Communication-efficient learning of deep networks from decentralized data,” *arXiv preprint arXiv:1602.05629*, 2016.
- [16] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, “A survey on mobile edge computing: The communication perspective,” *IEEE Communications Surveys & Tutorials*, vol. 19, no. 4, pp. 2322–2358, 2017.
- [17] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, “Federated learning: Challenges, methods, and future directions,” *arXiv preprint arXiv:1908.07873*, 2019.
- [18] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
- [19] K. Mishchenko, E. Gorbunov, M. Takáč, and P. Richtárik, “Distributed learning with compressed gradient differences,” *arXiv preprint arXiv:1901.09269*, 2019.
- [20] X. Liu, Y. Li, J. Tang, and M. Yan, “A double residual compression algorithm for efficient distributed learning,” *arXiv preprint arXiv:1910.07561*, 2019.
- [21] S. Vaswani, F. Bach, and M. Schmidt, “Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron,” *arXiv preprint arXiv:1810.07288*, 2018.
- [22] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, “Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization,” *arXiv preprint arXiv:1909.13014*, 2019.
- [23] Z. Li, D. Kovalev, X. Qian, and P. Richtárik, “Acceleration for compressed gradient descent in distributed and federated optimization,” *arXiv preprint arXiv:2002.11364*, 2020.
- [24] S. Khirirat, H. R. Feyzmahdavian, and M. Johansson, “Distributed learning with compressed gradients,” *arXiv preprint arXiv:1806.06573*, 2018.
- [25] H. Tang, X. Lian, C. Yu, T. Zhang, and J. Liu, “Doublesqueeze: Parallel stochastic gradient descent with double-pass error-compensated compression,” *arXiv preprint arXiv:1905.05957*, 2019.
- [26] P. Goyal, P. Dollár, R. Girshick, P. Noordhuis, L. Wesolowski, A. Kyrola, A. Tulloch, Y. Jia, and K. He, “Accurate, large minibatch sgd: Training imagenet in 1 hour,” *arXiv preprint arXiv:1706.02677*, 2017.
- [27] Y. You, Z. Zhang, C.-J. Hsieh, J. Demmel, and K. Keutzer, “Imagenet training in minutes,” in *Proceedings of the 47th International Conference on Parallel Processing*, p. 1, ACM, 2018.
- [28] M. Yamazaki, A. Kasagi, A. Tabuchi, T. Honda, M. Miwa, N. Fukumoto, T. Tabaru, A. Ike, and K. Nakashima, “Yet another accelerated sgd: Resnet-50 training on imagenet in 74.7 seconds,” *arXiv preprint arXiv:1903.12650*, 2019.
- [29] Z.-Q. Luo, “On the convergence of the lms algorithm with adaptive learning rate for linear feedforward networks,” *Neural Computation*, vol. 3, no. 2, pp. 226–245, 1991.
- [30] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” in *Advances in neural information processing systems*, pp. 315–323, 2013.
- [31] A. Defazio, F. Bach, and S. Lacoste-Julien, “Saga: A fast incremental gradient method with support for non-strongly convex composite objectives,” in *Advances in neural information processing systems*, pp. 1646–1654, 2014.

- [32] S. Shalev-Shwartz and T. Zhang, “Stochastic dual coordinate ascent methods for regularized loss minimization,” *Journal of Machine Learning Research*, vol. 14, no. Feb, pp. 567–599, 2013.
- [33] C. Liu and M. Belkin, “Mass: an accelerated stochastic method for over-parametrized learning,” *arXiv preprint arXiv:1810.13395*, 2018.
- [34] W. Y. B. Lim, N. C. Luong, D. T. Hoang, Y. Jiao, Y.-C. Liang, Q. Yang, D. Niyato, and C. Miao, “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, 2020.
- [35] F. Zhou and G. Cong, “On the convergence properties of a k -step averaging stochastic gradient descent algorithm for nonconvex optimization,” *arXiv preprint arXiv:1708.01012*, 2017.
- [36] S. U. Stich, “Local sgd converges fast and communicates little,” *arXiv preprint arXiv:1805.09767*, 2018.
- [37] J. Wang and G. Joshi, “Cooperative sgd: A unified framework for the design and analysis of communication-efficient sgd algorithms,” *arXiv preprint arXiv:1808.07576*, 2018.
- [38] A. Khaled, K. Mishchenko, and P. Richtárik, “First analysis of local gd on heterogeneous data,” *arXiv preprint arXiv:1909.04715*, 2019.
- [39] H. Yu, S. Yang, and S. Zhu, “Parallel restarted sgd with faster convergence and less communication: Demystifying why model averaging works for deep learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5693–5700, 2019.
- [40] B. E. Woodworth, J. Wang, A. Smith, B. McMahan, and N. Srebro, “Graph oracle models, lower bounds, and gaps for parallel stochastic optimization,” in *Advances in neural information processing systems*, pp. 8496–8506, 2018.
- [41] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, “On the convergence of fedavg on non-iid data,” *arXiv preprint arXiv:1907.02189*, 2019.
- [42] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [43] J. N. Tsitsiklis and Z.-Q. Luo, “Communication complexity of convex optimization,” *Journal of Complexity*, vol. 3, no. 3, pp. 231–243, 1987.
- [44] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th international conference on machine learning (ICML-10)*, pp. 807–814, 2010.
- [45] P. Elias, “Universal codeword sets and representations of the integers,” *IEEE transactions on information theory*, vol. 21, no. 2, pp. 194–203, 1975.